

中文電腦缺字解決方案

莊德明 中央研究院資訊科學研究所

derming@gate.sinica.edu.tw

摘 要

使用電腦來處理漢字資料時，常會遇到一些電腦沒有的字形，以致無法輸入，這便是缺字問題。面對缺字問題，雖然擴大電腦字集是該努力的工作，但是並不能徹底解決問題。我們認為，要解決缺字問題，還是應該在電腦中建構一個處理字形的機制。

本所自 1993 年起，即研擬漢字在計算機中的制式表達，稱為構字式，以登錄漢字的字形結構，並建置漢字構形資料庫。構字式係將漢字視為一個二度空間圖案，再由部件以橫連、直連、包含三個法則拼湊而成。構字式可用來表達缺字，同時也可以當作漢字的交換碼。

由於近年來網際網路蓬勃發展，資訊共享的需求日益殷切，本所也和「教育部圖書館自動化及網路化策略推動委員會」合作，推動以漢字構形資料庫來解決中文電腦的缺字問題，第一期計畫於 2000 年底結束，並且推出可在 Microsoft Windows 95 或 98 中文繁體版上使用的光碟。這片光碟已收錄《漢語大字典》的 54,678 個字形及 12,208 組異體字。光碟安裝後，使用者可以利用部件迅速檢字，並在 Microsoft Word 97 或 2000 中文繁體版上使用這些字形。

漢字構形資料庫的建置仍在進行中，第二期計畫從 2001 年 8 月到 2002 年 6 月，預計增收《說文解字》的 9353 個小篆字頭及 1163 個重文，以銜接古今文字。小篆字形的登錄，必須依據《說文解字》的釋形，以建立文字學資料庫。至於小篆和楷體間的對映，可以參考《漢語大字典》及《說文解字詁林》。部份小篆字可能對映好幾個楷體字，這部分可參考《漢語大字典》的異體字表。

小篆以外的古文字，如甲骨文、金文、楚簡等，本所也和院內的歷史語言研究所合作，目前以金文的進展最快，可望和小篆同時收錄在第二期的光碟中。

壹、前言

使用電腦來處理漢字資料時，常會遇到一些電腦沒有的字形，以致無法輸入，這便是缺字問題。尤其是在處理古代文獻時，缺字的情況特別嚴重。

為了應付缺字問題，一般的治標方法是在電腦字集的外字區中找一個碼位，造上所缺的字形，這種作法雖然可以在該電腦中顯示缺字，但也面臨了資訊共享的障礙及造字空間不足的問題。另外，擴大電腦字集也曾經是大家的努力方向，但是經過多年的驗證，仍然發現字形無法收齊，改寫應用程式的成本過高。我們認為要徹底解決缺字問題，在電腦中不能僅靠一個碼位來代表漢字，而應該在電腦中建構一套處理字形的機制。

本所自 1993 年起，即研擬漢字在計算機中的制式表達，以登錄漢字的字形結構，並建置漢字構形資料庫。近年來由於網際網路蓬勃發展，資訊共享的需求日益殷切，本所也和「教育部圖書館自動化及網路化策略推動委員會」合作，推動以漢字構形資料庫來解決中文電腦的缺字問題。第一期期畫於 2000 年底結束，並且推出可在 Microsoft Windows 95 或 98 中文繁體版上使用的光碟。這片光碟的主要內容如下：

1. 涵蓋《漢語大字典》[1]的 54,678 個字形及繁體中文 Big5 的 13,053 個字形，合計 54,711 個字形。
2. 收錄《漢語大字典》的異體字表，共 12,208 組，包含 36,309 個字形。
3. 提供 4,766 個部件以檢索字形，其中包含 1,324 個字根。1,324 個字根，可再分成常用字根 657 個，間用及罕用字根 667 個；常用字根可涵蓋 52773 個字形，間用及罕用的字根涵蓋 1938 個字形。
4. 擴充 Microsoft Word 97 及 2000 繁體中文版的功能，可在 Word 下使用 54,711 個漢字。
5. 開發漢字構形資料庫使用界面，讓使用者可以快速的檢索缺字及異體字。
6. 提供線上說明及使用手冊。

漢字構形資料庫的建置仍在進行中，第二期計畫從 2001 年 8 月到 2002 年 6 月，工作重點如下：

1. 將 Word 上的缺字處理功能移植到瀏覽器，如 Internet Explorer 或 Netscape，以解決網路的缺字問題。
2. 訂定字根的標準制式結構，以處理筆劃上的字形變化。
3. 增加收錄的字形。
4. 處理《說文解字》的 9353 個小篆及 1163 個重文，以銜接古今文字，並建立文字學資料庫。

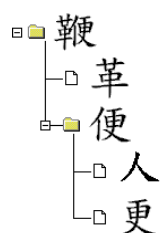
5. 提供使用者利用網際網路更新本系統，如資料庫及應用程式等。

至於小篆以外的古文字，如甲骨文、金文、楚簡等，本所也和院內的歷史語言研究所合作，目前以金文的進展最快，可望和小篆同時收錄在第二期的光碟中。

本著推廣漢字構形資料庫的立場，第一期的光碟將隨著本文發送，提供與會者使用，也希望大家提出批評，以作為下一版改進的依據。漢字構形資料庫的使用，可在光碟安裝完畢後，直接列印使用手冊或閱讀線上說明。與會者若想在安裝光碟之前，能對本系統的設計理念及操作方式有所了解，可直接參考本文。本文除了描述如何在電腦中表達缺字外，也會說明 Word 中的缺字處理，以及漢字構形資料庫使用界面的主要功能。另外，第二期計畫中有關古今文字的銜接機制，也將在本文中一併提出。

貳、缺字在電腦中的表達

有別於以往在電腦中僅靠一個碼位來識別漢字，在漢字構形資料庫中，每個漢字都必須登錄結構，並以構形來識別。有關漢字的構形分析，可參考《漢字漢語基礎》[2]一書，該書要點摘錄如下，字例可見圖一的「鞭」及圖二的「牖」。



圖一、「鞭」的構形分析














圖二、「牖」的構形分析



















1. 漢字的構形單位是部件，如「革」、「便」是「鞭」的部件。
2. 部件是有層次的，最小而不再拆分的基礎部件，稱為形素或字根，如「革」、「人」、「更」是「鞭」的字根。
3. 漢字是由有限的字根來組成數以萬計的字形。
4. 漢字的組合類型有層次結構及平面結構兩種。
5. 層次結構的字是由部件分作若干層次逐步累加而構成的，如「鞭」，漢字的結構大都屬於此類。
6. 平面結構的字是由三個以上的部件一次性組合，分不出層次來，如「牖」。
7. 在依層次拆分的漢字中，部件是有層級的，如「革」、「便」為「鞭」的一級部件，「人」、「更」為二級部件。一級部件也稱直接部件。

漢字構形資料庫的字形登錄，並不強調構形理據，而是將漢字視為一個二度空間圖案，再由部件拼湊而成。即使把漢字當作部件拼湊的圖案，上述的構形分

析大致上仍可沿用，只是有以下幾點不同：

1. 層次結構的字形在登錄時，必須標示部件間的連接方式；這些連接方式有橫連、直連及包含三種。如「鞭」由「革」橫連「便」，「童」由「立」直連「里」，「趙」由「走」包含「卓」。
2. 部件的登錄直接依照字樣的形體，如「攢」的部件「刀」登錄成「𠂇」，這樣才能識別「攏」及「攢」二字。
3. 層次結構的字形在登錄時，只須登錄直接部件，整個字形結構即可透過遞歸(recursive)的方式完成。如「鞭」為「革」橫連「便」，「便」為「亻」橫連「更」，而「革」、「亻」、「更」為字根。
4. 平面結構的字形，若是在電腦中找到合適的部件，即可登錄成層次結構。如「碧」的構形理據是由「王」、「白」、「石」三個部件一次性組合，但在 Big5 字集有「珀」字，所以「碧」即可登錄成「珀」直連「石」。
5. 層次結構的字形，若是在電腦中找不到合適的部件，也只好登錄成平面結構。如「瑩」的部件「憤」是 Big5 的缺字，假如不在 Big5 外字區補上「憤」字，「瑩」只好登錄成「巾」、「口」、「玉」三個部件一次性組合。
6. 無法以橫連、直連或包含方式連接的部件，一律視為字根，不再拆分，如「重」、「東」。

為了表達上的簡潔，我們用 、、 三個連接符號，分別表示橫連、直連及包含，這時「鞭」的字形結構表達成『鞭=革便』，其中『革便』稱為構字式。同理，『童=立里』、『趙=走卓』。構字式是用來表達層次結構的字形，至於平面結構的字形，可用部件序來表達。部件序是將字形的構成部件，以書寫順序排列；使用部件序時，必須在部件的前後加上起點標示及結束標示，如『牖=片戶甫』。

構字符號除了連接符號及起止標示外，還有方便符號。方便符號可用來表達相同部件的組合，包括二、三或四個相同部件橫連(, , )，二、三或四個相同部件直連(, , )，三個相同部件呈品狀排列()，四個相同部件呈罈狀排列()。方便符號可單獨使用，使用時必須放在部件的前面，如『囍=喜』、『棗=束』、『龔=龍』、『焱=火』；方便符號也可用在構字式或部件序中，如『壘=田土』、『蹉=𠂇大牛』。

第一期光碟收錄的 54,711 個漢字中，採用構字式表達的有 45,063 個，部件序表達的有 8,635 個，只用方便符號表達的有 317 個，其他的 696 個為字根。這些字形拆分後，可得到 4,766 個部件，其中字根佔 1,324 個。4,766 個部件中，有 3,391 個是 Big5 電腦字集用字，另外的 1,375 個則是新增的部件外字；新增的部件外字，字根就佔了 994 個。部件總數及外字如表一。

表一、部件總數及外字

	字根	非字根	合計
Big5	330	3,061	3,391
外字	994	381	1,375
合計	1,324	3,442	4,766

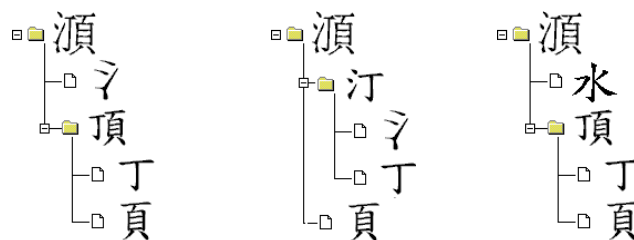
部分字根由於意義相近或形體相似，可歸納成異體字根，如「刀」和「刂」、
「日」和「曰」。目前初步整理出異體字根 296 組，共包含 716 個字根。透過異體
字根，無論是由部件檢字或表達構字式，都能提供較大的彈性。如部件「刀」可
同時找出「懾」及「攢」，構字式『畫_△刂』或『畫_△刀』都可對映「劃」。

在 1,324 個字根中，可再依使用頻次分為常用字根 657 個，間用字根 539 個，
罕用字根 128 個。常用字根可涵蓋 54,711 個字形中的 52773 個字形，間用字根涵
蓋 1810 個字形，罕用字根只涵蓋 128 個字形。

雖然在漢字構形資料庫中，每個字的表達式都不相同；但對使用者而言，同
一個字是可以有多種不同的表達式，舉例如下：

1. 不同的構字式，如『曾=彳_△頂』、『曾=汀_△頁』。
2. 構字式或部件序，如『琿=魚_△嵩』、『琿=魚_△崔岡』。
3. 構字式或方便符號，如『疍=魚_△〇〇魚』、『疍=魚_△魚』。
4. 不同的部件序，如『趟=彳_△西域哲』、『趟=彳_△西域哲人』。
5. 本是 Big5 用字，卻誤以為缺字，如『蠡=虫_△嵩』。
6. 異體字根，『曾=彳_△頂』、『曾=水_△頂』。

上述的表達式雖然不同，但是在經過異體字根轉換後，它們的字根都相同。
資料庫中只須登錄最精簡或最常用的一個表達式，其他的表達式可透過字根的
分析找到對映的字形。如圖三中「曾」的三個構字式，字根都是「水」、「丁」、「頁」，
資料庫只須登錄『彳_△頂』即可。



圖三、「曾」的構字式

漢字的結構表達式雖然有構字式、部件序及方便符號三種，但是為了陳述的
方便，以下若無特別聲明，「構字式」也同時涵蓋了部件序及方便符號，而視為漢
字結構表達式的同義詞。

使用漢字構形資料庫來解決缺字問題，所有的缺字都可採用構字式來表達，

這種表達式有以下的優點：

1. 構字式可用作字形的識別碼，因為每個漢字的構字式都不相同。
2. 構字式是用限的部件來表達數以萬計的漢字，因此構字式不佔碼位，不會有造字空間不足的問題。
3. 構字式能表達的漢字，遠比目前任何一個電腦字集還要多，因此構字式更適合充當漢字的交換碼，來突破資訊共享的障礙。
4. 構字式是一種知識表達，不僅較數字碼易讀易懂，所孕藏的構字知識更有利於應用程式的處理。
5. 構字式可及早協助使用者輸入缺字，減少資料登錄的成本。

構字式雖然有這些優點，但是由於缺字的表達和原先的電腦字集不同，所以諸如字形的顯示、檢索等相關的應用程式都需要修正，或者在這些應用程式前後增補前置或後續處理程序。但是和更換電腦字集比起來，這些程式的修正或程序的增補成本仍然低得多。

參、銜接古今文字

漢字構形資料庫的第二期計畫，預計增收《說文解字》中的 9353 個小篆及 1163 個重文，以銜接古今文字，並建立文字學資料庫。另外，本所也和院內的歷史語言研究所合作，將金文納入本系統，以便和小篆同時收錄在第二期的光碟中。這個工作還在進行中，本節只以小篆為例，說明古今文字銜接的機制。

目前電腦字集通常採用楷書作為標體字樣，若要使用小篆等古文字，通常是透過另一個字型檔，在對映的楷體字的碼位上造字。小篆和楷體的對映，可參考《說文解字詁林》[3]及《漢語大字典》，字例可見表二。

表二、小篆和楷體的對映字例(Big5)

Big5	BBCE	B1B5	A56A	A4B5	A4E5	A672
楷體	銜	接	古	今	文	字
小篆	銜	接	古	今	文	字

然而這種作法，同樣會有缺字問題，缺字的來源有以下幾種，字例可見表三：

表三、小篆和楷體的對映字例(缺字)

Big5	AAB1	(缺)	A7F3	(缺)	(缺)
構字式	王𠃉元	貝𠃉元	更	丙𠃉支	行𠃉彡
楷體	玩	爾	更	惓	卷
小篆	玩	𠃉	更		𠃉

1. 重文。如「𠃉」為「玩」的或體，兩者對映的楷體字分別為「爾」及「玩」，

其中「繭」為 Big5 缺字。

2. 楷化字。如「更」隸定為「更」，楷化成「愄」，其中「愄」為 Big5 缺字。這些楷化字可參考《漢語大字典》的異體字表，如圖四。



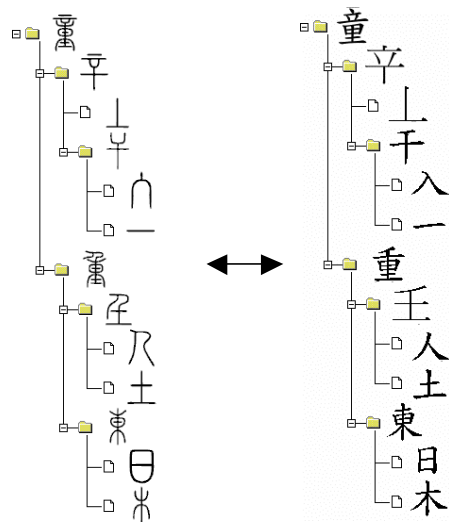
圖四、「更」的異體字

3. 罕用字。如「𪛗」對映的楷體字「卷」。

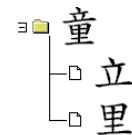
從表三可得知，由於構字式所能表達的漢字，遠比目前任一個電腦字集還多，連帶的也讓小篆等古文字在使用上更具彈性。

雖然小篆是透過對映的楷體字來使用，但在字形結構的登錄，則和楷體字不同：

1. 楷體字的登錄是將漢字視為一個二度空間圖案，再由部件拼湊而成；小篆的登錄，必須參照《說文解字》的釋形。圖五為「童」字結構及對映的楷體字，「童」字結構則見圖六。使用部件檢字，「童」可由「立」、「里」來檢索，而「童」可由「辛」、「重」、「上」、「干」等部件來檢索。



圖五、「童」字結構及對映的楷體字



圖六、「童」字結構

2. 小篆部件的登錄必須增加表形、聲、或省形的標示。如「童」的部件「辛」表形，「重」表省聲。

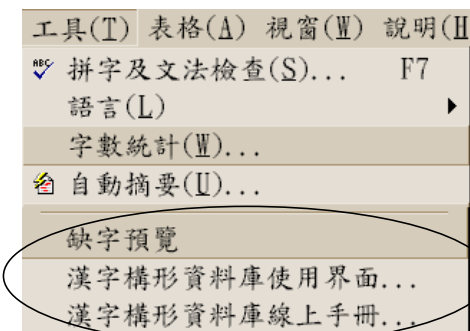
由小篆的例子可以看出，在古今文字銜接的機制上，古今文字的字形必須分別登錄，然後再建立字形間的對映。

肆、在 Word 中處理缺字

在 Microsoft Windows 下，Microsoft Office 不但是使用最廣泛的應用軟體，同時還可充當應用軟體的開發平台。在第一期的計畫光碟中，我們提供了 Word 的缺字處理範本，這個範本包含「構字符號」工具列及缺字處理次功能表。



圖七、構字符號工具列

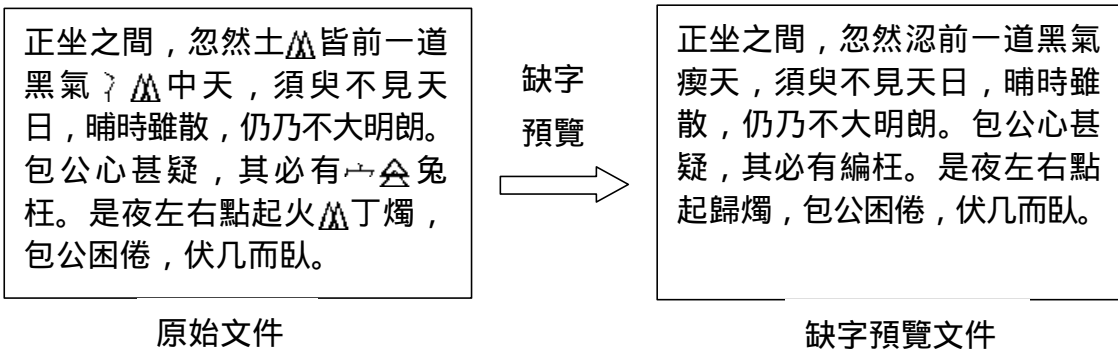


圖八、缺字次功能表

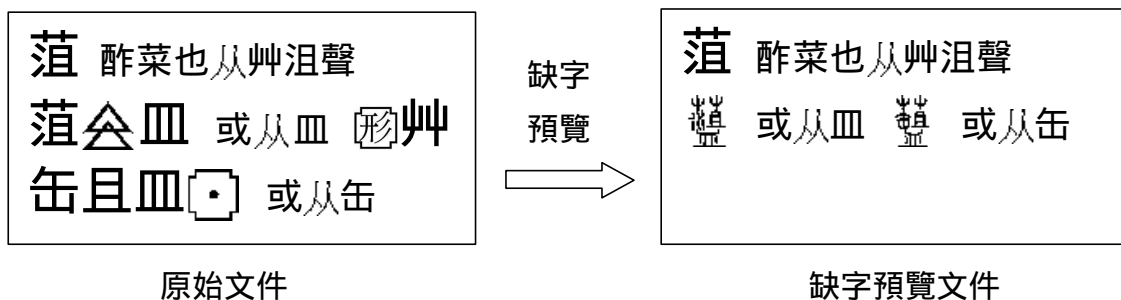
在 WORD 中缺字的輸入，是以構字式來表達。使用者輸入構字式的文件，稱為「原始文件」。執行缺字預覽，可將構字式轉成字形，並存於新增的文件中，這個新增的文件稱為「缺字預覽文件」，如圖九、圖十。由於構字式可當交換碼，所以原始文件適合交換及修改。缺字預覽文件只供閱讀及列印，修改時必須回到原始文件進行。

「構字符號」工具列如圖七，可協助使用者輸入構字符號。

缺字處理次功能表如圖八，它是位於「工具」功能表之下，由「缺字預覽」、「開啟漢字構形資料庫使用界面」、「閱讀漢字構形資料庫線上手冊」三個次功能表組成。缺字預覽可將文件中的構字式轉成適合閱讀及列印的字形；漢字構形資料庫使用界面除了可協助使用者輸入缺字外，其他的重要功能可參考下節；至於詳細的操作程序可閱讀線上手冊。

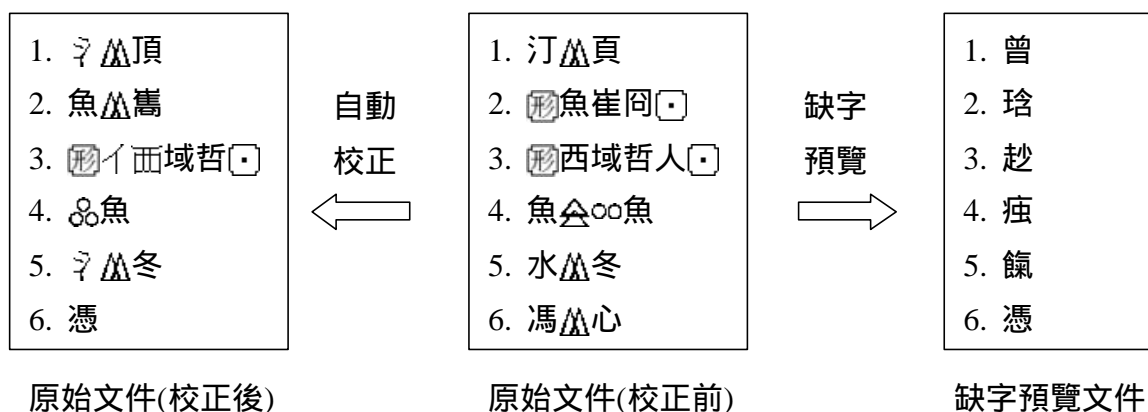


圖九、缺字預覽(一)



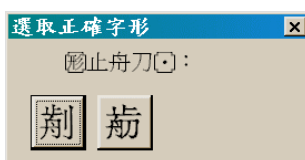
圖十、缺字預覽(二)

一般而言，缺字預覽並不會更動原始文件的內容；但是由於同一個缺字，使用者可能採用不同的構字式，以致和資料庫登錄的不同，這時缺字預覽會進行自動校正。自動校正會將原始文件中的構字式改成資料庫所登錄的，程序如圖十一。



圖十一、自動校正(一)

自動校正時，若是發現構字式對映的字形不只一個，會列出可能的字形，要求使用者確認，再進行校正。如「荖」、「臄」登錄的構字式分別為『止𠂔𠂔』、『𠂔止舟刀』，假使輸入的構字式為『𠂔止舟刀』時，由於無法在資料庫中找到相同的構字式，而且經過異體字根替換後，發現「荖」、「臄」兩個字形都有可能，於是要求使用者確認，如圖十二。

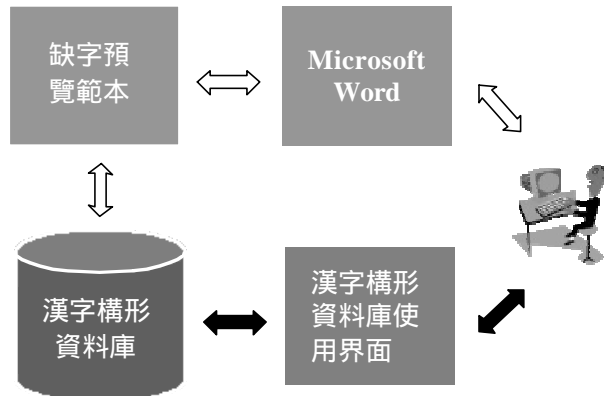


圖十二、自動校正(二)

由於 Word 不單是純文字編輯器，目前除了正文可執行缺字預覽外，還包括註腳、文字方塊、頁首、頁尾及註解。本論文即是應用缺字預覽的一個實例。

伍、漢字構形資料庫使用界面

第一期的光碟片中包含了兩個漢字構形資料庫的應用程式，一個是在 Word 的缺字預覽，另一個則是「漢字構形資料庫使用界面」，系統架構見圖十三。對大多數使用者而言，他們比較關心的是如何在 Word 中使用更多的字形；但是要了解漢字構形資料庫的全貌，還是要從使用界面著手。本節介紹的部分操作畫面係後來新增，已和所附的光碟不同。



圖十三、漢字構形資料庫的應用

漢字構形資料庫目前收錄的字形包含 54,711 個楷體字，這些楷體字可再分成常用字(3915)、五大碼(13,053)、簡化字(2,235)、及《漢語大字典》(54,678)四個字集；新增收的 9353 個小篆及 1163 個重文，則以《說文解字詁林》為底本。這些字集可參考圖十四的字集功能表。

漢字構形資料庫使用界面的主要功能可參考圖十五的字形功能表，包含字形孳乳、字形結構、字形演變、字形索引及異體字表查詢。

字集
常用字
五大碼(Big5)
簡化字總表
✓ 漢語大字典
說文解字詁林

圖十四、字集功能表

字形
字形孳乳...
字形結構...
字形演變...
字形索引...
異體字表...

圖十五、字形功能表

字形孳乳是透過部件來檢索字形，和傳統字典的部首檢字相比，部件檢字的特點如下：

1. 部件的個數遠比部首來得多。如《康熙字典》的部首有 214 個，《說文解

字》的部首有 540 個，然而目前收錄的部件有 4,766 個。

2. 構成字形的任一個部件都可用來檢索。如採用部首檢字，「算」必須查「口」部；但採用部件檢字，「算」可由任一個部件「口」、「金」、「欠」、「欽」查得。
3. 可利用多個部件同時檢索，迅速查得字形。表四為不同部件由《漢語大字典》檢索出的字形數，從結果可以看出，部件的層級越高、越不常用或越多，檢索的效果就越好，如「欽」、「欠」或「口金欠」。

表四、不同的部件檢索出的字形數

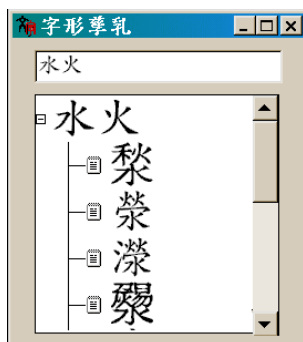
部件	口	金	欠	欽	口金	口欠	金欠	口欽	口金欠
字形數	7,887	1214	628	11	158	94	21	3	3

4. 配合異體字根可增加檢索的彈性。如「水」的異體字根有「氵」、「氷」，「火」的異體字根有「灬」、「灺」，由《漢語大字典》檢索包含部件「水」、「火」的字形，若是再加上異體字根，檢索出的字形會由原先的 6 個增加到 82 個，詳見表五。

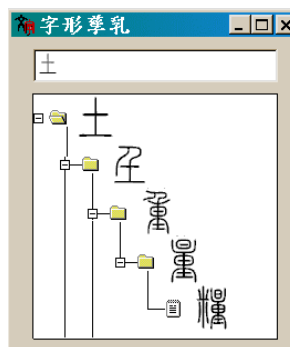
表五、配合異體字根檢索出的字形數

部件	水火	水灺	水灺	氵火	氵灺	氵灺	氷火	氷灺	氷灺
字形數	7	2	0	42	28	2	1	1	0
字形總數	82(「灺」字重複計算)								

部件檢字前，除了可先設定字集外，檢索結果也可分為一次列出或是層級列出。圖十六係由《漢語大字典》檢索包含部件「水」、「火」的字形(不含異體字根)，並且一次列出檢索結果。圖十七係由《說文解字詁林》檢索包含部件「土」的字形，但是改採層次列出。

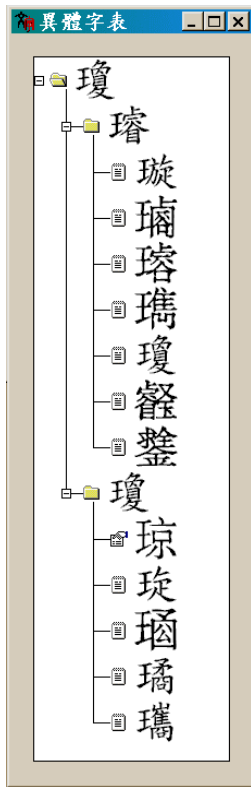


圖十六、字形孳乳(一)

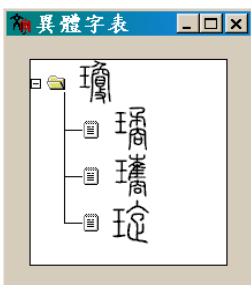


圖十七、字形孳乳(二)

在圖十六、圖十七的字形孳乳視窗中，只要用滑鼠點選檢索出的任一個字，該字的構字式就會自動複製到 Windows 的剪貼簿，此時即可在 Word 等應用程式貼上構字式以輸入缺字。除了複製構字式外，該字的異體字、字形結構、字形演




圖十八、「瓊」的異體字



圖十九、「瓊」的重文

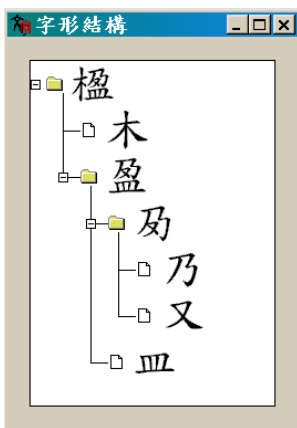
變及字形索引也會同時出現在相關的視窗。

圖十八，圖十九的異體字表視窗分別列出「瓊」的異體字及「瓊」的重文。「瓊」的異體字係節錄自《漢語大字典》的異體字表，「瓊」的重文可參考《說文解字》。

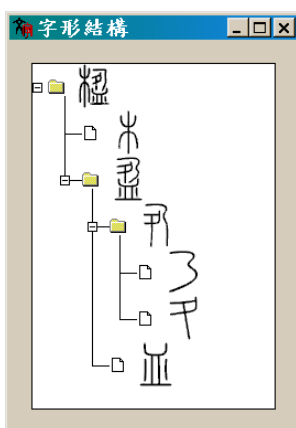
《漢語大字典》異體字表的編排是採用主體字統領異體字的方式，將同一主體字統領的簡化字、古今字、全同異體字（指音義全同而形體不同的字）非全同異體字（指音義部分相同的異體字），集中在該主體字下編為一組。這些異體字只有簡化字明確標示出，其餘並未說明。如圖十八「瓊」的簡化字「璿」前的。小篆的重文則有或體、俗字、奇字、篆文、古文及籀文，詳見《說文解字》。

異體字表中的 36,309 個字形，有 4,311 個出現一次以上。這些重複出現的字形，如「瓊」本身是主體字，還同時是「璿」的異體字。若是根據異體字表的部首、筆畫查閱，只能查出「瓊」為主體字；但在漢字構形資料庫使用界面中，不但列出「瓊」的異體字，還可同時查得「瓊」也是「璿」的異體字，並列出「璿」的所有異體字。

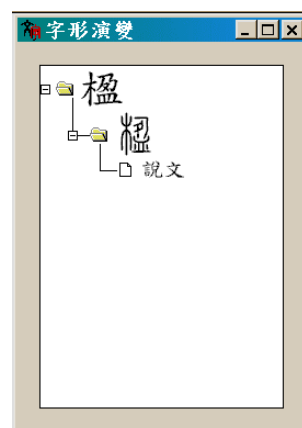
圖二十、圖二十一的字形結構視窗分別列出「楹」及「楹」的字形結構。圖二十二的字形演變視窗則為「楹」的字形演變，目前只列出小篆，往後還會加入金文。



圖二十、「楹」的字形結構

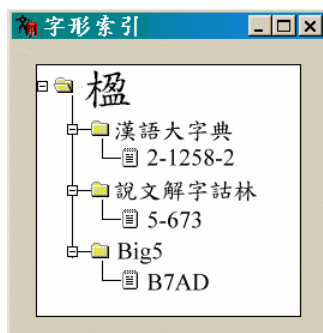


圖二十一、「楹」的字形結構



圖二十二、「楹」的字形演變

圖二十三的字形索引視窗列出「楹」在《漢語大字典》、《說文解字詁林》及 Big5 的索引，包括冊-頁-字、內碼。



圖二十三、「楹」的字形索引

陸、結語

本所建置漢字構形資料庫已進入第八年，大多數的時間都在整理基本資料，而無暇從事應用軟體的開發，以致研究成果一直無法推廣。這一次推出的光碟，雖然只能在 Microsoft Window 繁體中文版上搭配 Word 使用，但已算是一大進展。漢字構形資料庫的建置，包括基本資料的整理及應用軟體的開發，仍有很大的發展空間，這些絕非本所可獨立完成，所以本所也希望能和院內或院外的相關單位合作，以縮短研發時間，並將成果和大家一起分享。

參考文獻

1. 《漢語大字典》，徐中舒主編，遠東圖書公司，1991年9月。
2. 《漢字漢語基礎》，王寧主編，北京科學出版社，1996年7月。
3. 《說文解字詁林》，楊家駱主編，鼎文書局，1994年3月。